John von Neumann Institute for Computing


# The E-Cell Project and Challenges in Computational Systems Biology

K. Takahashi

published in

*From Computational Biophysics to Systems Biology (CBSB07)*,
Proceedings of the NIC Workshop 2007,
Ulrich H. E. Hansmann, Jan Meinke, Sandipan Mohanty,
Olav Zimmermann (Editors),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **36**, ISBN 978-3-9810843-2-0, pp. 55-60, 2007.



http://www.fz-juelich.de/nic-series/volume36

# The E-Cell Project and Challenges in Computational Systems Biology

**Koichi Takahashi**[1,2]

[1] The Molecular Sciences Institute,
2168 Shattuck Avenue, Berkeley, CA 94704, USA.
*E-mail: ktakahashi@molsci.org*

[2] Institute for Advanced Biosciences,
Keio University, Fujisawa, 252-8520, Japan.

Summarizing our experience in launching and running the E-Cell Project over the past ten years, I will discuss some of major challenges we believe we will face in the next ten years of cell and systems biological simulation, including the following two; (1) Undeniably the last ten years of computational systems biology has been (re)discovery of the biggest bottleneck in biochemical modeling; the lack of high-throughput and reliable means of obtaining reaction rate coefficients. Computational aids in determination of reaction rate coefficients will be one area in which fruitful interactions between molecular biology, biophysical chemistry, and supercomputing are highly expected. (2) Macromolecular crowding is ubiquitous and found in all types of cellular organisms on the earth, which can, when coupled with localization and diffusion, alter biochemical dynamics, change equilibrium points, slow down and change the manner how big molecules diffuse, and amplify intrinsic noise. It is also a suspected physico-chemical factor behind the emergence of eukaryotic organisms. Development of formal treatment and computational methods for crowded intracellular media will be some of the most important tasks left for computational biologists.

## 1 Introduction

The E-Cell Project [a] was started in 1996 with the aim of establishing technological bases that will make possible predictive modeling and simulation of cellular systems at the molecular level. Although we are halfway towards the ultimate goal we defined 11 years ago, I believe we are now at a good position where we can speculate, and, to some extent, predict the next 10 years of cell simulation technology by extrapolating our experience in the last decade, with the hope that it will be of any help in thinking about challenges of what nature we will encounter and what will have been done if we would be overcoming these. But before that, in the next section, let us think about why we want to model biochemical and cellular systems, and what we could do in the first place. I will then briefly look back the history of the E-Cell Project and our simulation platform E-Cell System, and discuss some challenges I believe we will face in the process of establishing cell simulation technology.

## 2 Simulation and Science

Why do we want to model and simulate biological systems? Why do we want to do it at the molecular level, and what insight do we expect from it?

[a] `http://e-cell.org`

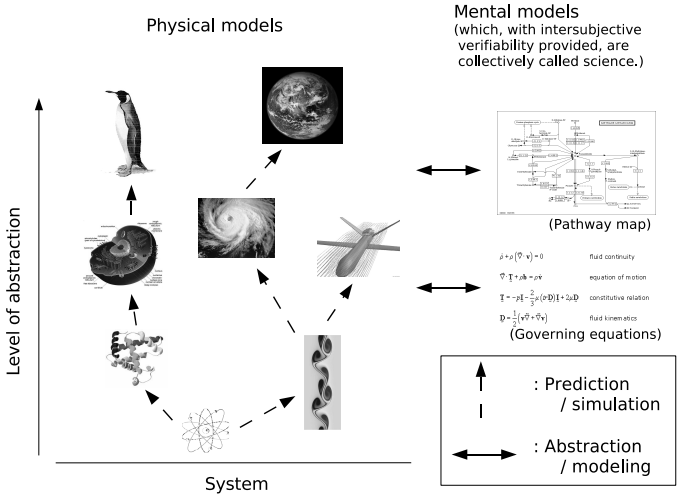


Figure 1. Physical models, mental models, and how these models are obtained or related by means of predictive and abstraction powers of modeling and simulation.

Science is a collection of mutually interconnected mental models that are intersubjectively verifiable and shared by the community of scientists. As a first approximation, we could define scientific research as forming of such mental models. In Figure 1, on the right hand side are two examples of such models, a metabolic pathway map and a set of equations for fluid dynamics. Time-space reproduction of molecules constituting the physical systems, such as those shown on the left hand side of the figure, do not by themselves represent understanding of how the systems work. Such physical models must somehow undergo a process of abstraction to construct models that have the power to predict and/or to explain behaviors of systems in different states or of different variations of the same type. Simulation allows us to leap forward exactly one step in the level of abstraction, from lower to higher. For example, equations about inter-atomic forces could be extrapolated to depict how proteins fold through the medium of numerical simulation. Or, atmospheric data gathered at a geographically sparse set of observation points for a certain period of time may lead to a model of a larger area, longer time-scales, and/or finer resolutions. Let us call this ability of simulation that enables us to interpolate or extrapolate physical models at certain levels of abstraction beyond our accessibility to the real-world target systems a *predictive power* of numerical simulation.

E-Cell Project is interested in modeling and simulating cellular systems at the level of the system-level output, or, physiology. If we want to build physiological models of the cell equipped with the predictive power, we have to model the system at one step lower in the hierarchy of abstraction. How big or small this 'one step' should be is subject to many technological factors. In principle, modeling becomes easier as we go lower in the abstraction level, since composite phenomena and complex components in the system tend to be decomposed to more basic principles and simpler components that are easier to model, and in many cases, less complicated to experimentally measure necessary parameters. This is

a crucially important aspect of the picture because the hardest part of cellular modeling and simulation lies in its ontological complexity, rather than emerging complexities seen in many other physical simulations[2]. On the other hand, how far we can go down given a desired observation frame highly depends on computational capacities and accuracy of numerical methods.

## 3 The E-Cell Project

What we aimed at when we started the E-Cell Project was to bring this predictive ability of numerical simulation into molecular biology. Our initial aim was to show it is not impossible to computationally reconstruct inner workings of cellular organisms at the whole cell-scale, and by doing so, to open a pathway that leads to establishment of cell simulation as a scientific method and as an inter-disciplinary technology that spans from measurement methods to numerical analysis.

We started the E-Cell project stimulated by the determination of the smallest known 580kb genome sequence of *Mycoplasma genitalium*. By the summer of 1997, we developed an early version of our modeling and simulation platform, the E-Cell System, and a model of a virtual cell of which 127 genes constitute its basic functions to sustain itself as a living organism; enzymes in the energy metabolism and phospholipid synthesis, and gene transcription and translation systems[1]. Since then, the E-Cell Project has been expanding areas of work to construct more detailed and precise models of cellular functions, such as the bacterial chemotaxis signaling pathway of *Escherichia coli* and metabolic pathways of human erythrocytes and mitochondria. Concomitant to the biological modeling projects was the development of the computational platform, the E-Cell System. We identified seven desirable features of cell simulation platforms to be truly useful, which constitute a quite different set of requirements than that of conventional physical simulators, such as the need for an integrative multi-algorithm, multi-timescale framework, object-orientation, dynamic model structure, and real-time user interactions. Due to the length limitation, interested readers are referred to other E-Cell Project publications[2] for more discussions. Since the development of the first version[1], our current version is E-Cell System version 3[3], which meets five of the seven requirements we defined. We have started the development of a new generation simulation kernel that will be the core of the E-Cell System version 4 to address the two remaining features, multi-spatial representations and dynamic model structure, in addition to a vastly improved support for parallel computation [b].

## 4 Some Important Challenges in Computational Systems Biology

Numerical simulation, as well as experimental measurement technologies, is placed at the center of disciplines playing important roles in computational systems biology, which is inter-disciplinary by nature. Here I pick up two (among many) major challenges in computational systems biology that we believe we will face in the course of pursuing cell simulation technology development.

[b] `http://e-cell.org/developers/e-cell-4`

### 4.1 Methods for Determination of Reaction Rate Coefficients

The last ten years have been rediscovery of the lack of high-throughput and reliable means of obtaining reaction rate coefficients, and this lack formed the biggest bottleneck in the biochemical modeling and simulation workflow. Many modeling projects got stuck as soon as they faced this lack of input parameters, and a common feature of the limited number of successful modeling projects has been the accumulation of a large body of kinetic studies for decades, each determined parameter for a particular enzyme corresponding to someone's doctoral degree.

When we think about these systems, however, we notice that there is no such measurable physical quantities like 'net reaction rate constants', but there is only coupling of two distinct physical occurrences; diffusive encounters of reactants and subsequent interactions between them. Let us think about bimolecular reactions, that are most commonly seen in cellular metabolic, signaling and gene expression pathways. Such reactions can formally be written as $A + B -> C$. Recall Arrhenius relation

$$k_{net} = A \exp\left(-\frac{E_a}{k_B T}\right), \tag{1}$$

where $k_{net}$ is the net reaction rate, $A$ is the frequency factor, $E_a$ is the activation energy, $k_B$ is the Boltzmann constant and $T$ is temperature. It can be read that there are two factors that determine the net rates of reactions, the frequency of molecular collisions and the activation energy which is related to interactions between molecules. Let us then look at this from another point of view;

$$\frac{1}{k_{net}} = \frac{1}{4\pi D\sigma} + \frac{1}{k_{intrinsic}}, \tag{2}$$

where $D$ is the diffusion coefficient, $\sigma$ is the effective cross section, and $k_{intrinsic}$ is the intrinsic reaction rate. When both of $A$ and $B$ molecules diffuse, $D$ becomes a relative diffusion coefficient. The first term in the right hand side indicates essentially the same as $A$, the frequency of collision, and the intrinsic reaction rate in the second term corresponds to the activation energy. Where did this $4\pi D\sigma$ come from? It came from the Smoluchowski theory of diffusion-limited reactions[7]. In the simplest case where we describe the system in one dimensional space,

$$k(t) = 4\pi\sigma^2 D \frac{\partial C}{\partial r} \tag{3}$$

where $t$ is the time since one of the reactive species was injected to the media while the other species was in a well-mixed state, $r$ is the distance between reactants, and $C$ is the concentration of the reactant. At the equilibrium ($t \to \infty$), and in the low-density limit, we get; $\lim_{t\to\infty} k(t) = \lim_{t\to\infty} 4\pi\sigma D \left(1 + \frac{\sigma}{\sqrt{\pi D t}}\right) = 4\pi\sigma D$. While this result partially applies to more complex systems[8], when concentration gradients and noise caused by the slow diffusion of macromolecules become relevant, or when density of molecules is high (e.g. see next section), simulation is typically the only way to accurately investigate the system. However, this decomposition of net reaction rates into two simpler, physically better formulated components may be one possible approach for cell simulation to overcome the bottleneck of reaction rates. An array of experimental methods that can be used to see how big molecules in the cell diffuse, such as fluorescence correlation spectroscopy and

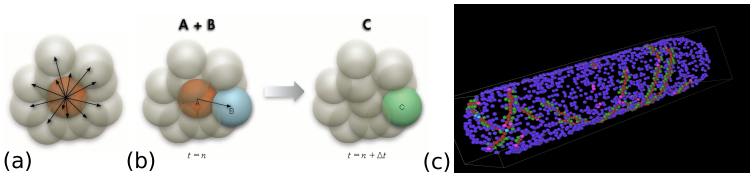


Figure 2. An extended lattice gas automata-like simulation method. (a) diffusion of a molecule to one of its 12 neighboring spheres in cubic-close packing. (b) a Monte-Carlo bi-molecular reaction $A + B \rightarrow C$ within the method with the time step $\Delta t$. (c) A sample simulation snapshot of MinD polymerization in *Escherichia coli* membrane. Reproduced with permission by Satya Arjunan[9].

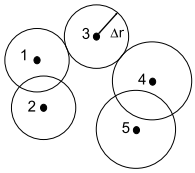


Figure 3. Basic idea used in the Green's Function Reaction Dynamics. Particles are 'protected' by protecting domains (spheres in this figure) to limit interactions up to two bodies. In this figure, the particles 1, 2, and 4, 5 form pairs, and the particle 3 remains single. Particles are propagated according to corresponding Green's functions for diffusion.

single particle tracking, are rapidly developing. At the same time, dynamics of protein-protein, protein-ligand, and protein-nucleic acid interactions are some of the areas where expected developments in high-performance scientific computing would be found most useful, ultimately leading to a high-throughput method of determination of reaction rates, possibly coupled with some form of high-sensitivity biosensors.

### 4.2 Macromolecular Crowding

Another area where supercomputing may stimulate scientific breakthroughs may consist in the striking nature of intracellular media. Extremely high densities of macromolecules (50-400 mg/ml, compare to 1-10 mg/ml typical *in vitro* conditions), called intracellular macromolecular crowding, is ubiquitous and found in all types of cellular organisms on the earth. Such non-ideal properties of the space in the cell result in different equilibrium points, altered reaction rates, slow and anomalous diffusion of macromolecules, and thus modified overall behaviors and dynamical characteristics of biochemical systems. Recently it was proposed that macromolecular crowding might be the physico-chemical culprit behind the emergence of eukaryotic cells[4].

Computational methods that could be used to study diffusion, localization and crowding of proteins in signaling pathways are reviewed somewhere else[5]. Here I mention two approaches the E-Cell Project is currently interested in. First of such approach belongs to a class of methods called cellular automata. Shown in Figure 2 is an extended lattice-gas automata. This class of method discretize the intracellular space with regular lattice, and propagate molecules to neighboring sites using Monte-Carlo calculations. When lattice

resolution is set adequately, the methods can give an approximate reproduction of molecular crowding. Lattice-based methods have a great affinity to modern digital computer architectures, and have promising scalability to supercomputers. Another class of method we are working to implement on the E-Cell System makes use of particles in continuum space and time[6]. This method called Green's Function Reaction Dynamics (GFRD) decomposes the multi-body problem that constitutes the biochemical system into a set of one and two-body problems (Figure 3), and corresponding Green's functions for diffusion-reaction are used to propagate the particles. GFRD is an accelerated form of Brownian Dynamics (BD) simulation method that is capable of effectively represent crowded space, and is typically up to five orders of magnitude faster than traditional approaches.

## 5 Conclusion

Understanding cellular systems as systems, not simply as a collection of molecular components, yet at the molecular level, is undoubtedly one of the most important scientific challenges in the 21st century, where computation may be the key to breakthroughs.

## Acknowledgments

I thank Satya N.V. Arjunan for the permission to use figures from his work, and Nathan Addy for his help in editing. I am a Human Frontier Science Program Fellow. The E-Cell Project is supported by CREST/JST and the MEXT of Japan (Leading Project for Biosimulation and the 21st Century COE Program).

## References

1. E-CELL: software environment for whole-cell simulation, M. Tomita, K. Hashimoto, K. Takahashi, T. S. Shimizu, Y. Matsuzaki, F. Miyoshi, K. Saito, S. Tanida, K Yugi, J. C. Venter, and C. Hutchison, Bioinformatics **15**, 1, 1999.
2. Multi-algorithm and multi-timescale cell biological simulation, K. Takahashi, PhD Thesis, Keio University (2004).
3. Multi-algorithm, multi-timescale method for cell simulation, K. Takahashi, K. Kaizu, B. Hu, and M. Tomita, Bioinformatics **20**, 4, 1999.
4. Genomics and the Irreducible Nature of Eukaryote Cells, C. G. Kurland, L. J. Collins, and D. Penny, Science **312**, 5776, 2006.
5. Space in systems biology of signaling pathways – intracellular molecular crowding in silico, K. Takahashi, S. Arjunan, and M. Tomita, FEBS Letters **579**, 8, 2005.
6. Green's-function reaction dynamics: a particle-based approach for simulating biochemical networks in time and space, J. S. van Zon and P. R. ten Wolde, J. Chem. Phys. **123**, 234910, 2005.
7. Diffusion-Limited Reactions, S. A. Rice, in Compr. Chem. Kinet., Vol. **25**, editec by C. H. Bamford, C.F.H. Tipper, R.G. Compton, Elsevier, New York, (1985).
8. Theory of reversible diffusion-influenced reactions, N. Agmon and A. Szabo, J. Chem. Phys. **92**, 5270, 1990.
9. A 3D pole-to-pole oscillation model of MinD on Escherichia coli membrane, S. N. V. Arjunan and M. Tomita, unpublished (2007).